Expertise of the Machine Learning Reading Group members

Nov 09th 2009.

Auto-tunning

by Seungjai Min

Programming heterogeneous systems, such as cluster of CPUs and GPUs is difficult and writing performance portable applications for such systems is even harder. For example, optimizing inter-processor communication strategy and searching for optimal problem partitioning (CPU/GPU, Process/Threads) to reduce load imbalance, are increasingly problematic with larger parallelism. Auto-tuning can be a right direction to these problems. However, auto-tuning algorithms face the challenge of exploring large parameter space. Thus, it is critical to develop an efficient search strategy - especially for runtime adaptive auto-tuning. We will study if SML can give us the timely solution from large quantities of data.

Computer vision / pattern recognition

by Daniela Ushizima

There are 4 main topics contemplating a joint effort between Visualization and Math groups: pattern recognition from medical images, artistic media and particle simulations. The machine learning algorithms under investigation include both non-supervised techniques, based on clustering (hierarchical, partitioning methods and model based) and supervised (Bayesian classifiers, SVM, neural networks, etc...) and combinations of the two.

On medical images, the goal is to characterize ocular fundus images to enable diagnosis and follow-up on eye diseases (retinopathy, glaucoma, laser treatment). This includes image segmentation and feature extraction using mathematical morphology and wavelets to identify vasculature system, exudates, microaneurysms, optical disc, macula and laser treated area. The final step is the automatic classification of different eye structures and abnormalities.

On synthetic aperture radar images (SAR), the goal is to identify objects from cluttered digital images. We have designed filters to remove speckle noise, combining local statistics with mathematical morphology. In addition, we proposed a new algorithm to segment point targets using front propagation, including speckle noise statistics to the numerical scheme.

On artistic media, the goal is to identify patterns on large image databases (paintings, logos, games, etc.), using low level image descriptors (intensities, texture, border) and PCA to combine such descriptors, for later clustering and visualization of the samples using distance similarities.

On particle acceleration, the goal is to track bunches of electrons under relativistic acceleration among millions of particles in a time series, generated by numerical simulations. Recent accomplishments include the use of sampling conditioned to the kernel density estimates. This idea enables grouping particles using mixture-model clustering, allied to the Bayesian Information Criteria to identify the best number of clusters to contain compact groups of particles.

Sparse Eigenvalue Problems in Nuclear Physics

by Christopher P. Calderon

This project is concerned with developing new computational approaches to perform eigen computations to better understand the forces governing particles making up the nucleus of different atoms. The matrices encountered are often sparse, but the sparsity pattern is not known a priori. Selecting a good basis (representation) of the underlying Hamiltonian can be very important in the design of fast and reliable algorithms. Efficient iterative methods are important in problems of the size we want to consider, so tools like MPI and OpenMP will be utilized to perform fast matrix vector products.

Many applications in manifold learning (such as local linear embedding) and sparse covariance matrix estimation can potentially benefit from the methods developed for this nuclear physics project. These types of extensions will also benefit the project described below.

Learning from Single-Molecule Fluctuations Using Functional Data Analysis

by Christopher P. Calderon

Recent advances in single-molecule physics are allowing researchers to experimentally track individual molecular scale events with spatial and temporal resolution that would have been considered nearly impossible just a decade ago. For example, it is possible to monitor the transcription of a single DNA into RNA. When one looks at these scales, the time series coming from the experiments contain inherent thermal fluctuations (noise) from multiple sources, e.g. solvent bombardment, molecular vibration, etc. These signals also contain instrument noise associated with the measuring apparatus. The thermal fluctuations are often physically relevant to biological function, but this noise is not uniform in state space and their properties are typically unknown from first principles considerations in systems studied at the single-molecule level.

A substantial challenge here is in summarizing the physically relevant information contained in the large amount of noisy time correlated data measured from a single experimental trajectory. It has been demonstrated that by fitting stochastic differential equations (SDEs) to the observed response, that this can provide a useful summary of the dynamics (and at the same time avoid excessive coarse graining). This summary transforms the rough trajectory (path) data into smooth functions, e.g. an effective force and local diffusion coefficient. Exploring computational approaches that pool multiple data sets together for both estimation (using mixed model regression and penalized spline smoothing techniques) and unsupervised learning (e.g. identify clusters of functions) are of interest here. Recent ideas from functional data analysis will likely assist such tasks.